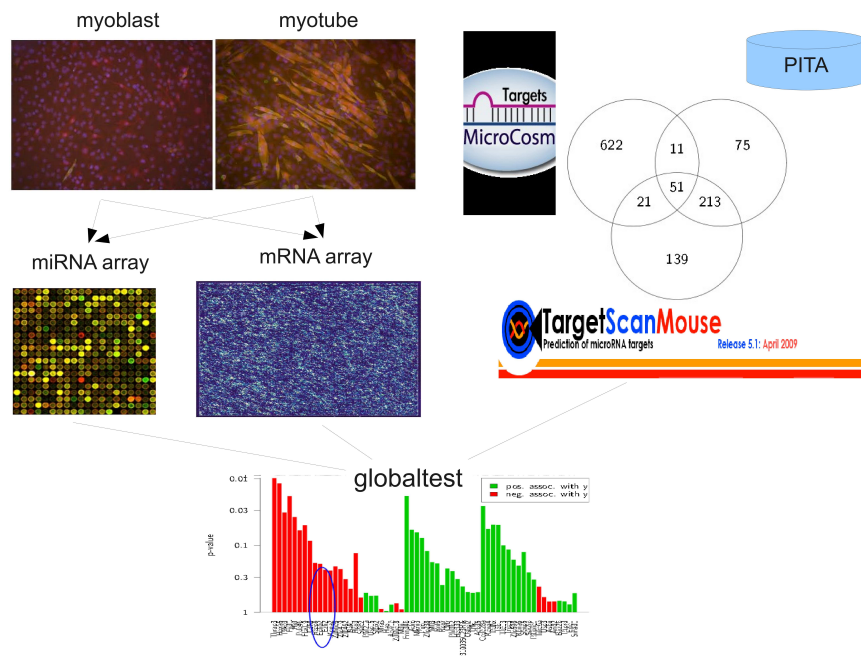


# 1 Scheme Integrated approach



*Integrated analysis of DNA copy number and gene expression microarray data using gene sets.*  
Menezes *et al.*, 2009, BMC Bioinformatics.

## 2 Statistical model

The global-test: an empirical Bayesian approach for gene set testing

$$Y_i = \alpha + \sum_{j=1}^m x_{ij} \beta_j + \epsilon_i \quad i = 1, \dots, n \quad (1)$$

$Y_i$ : expression of a microRNA on  $i^{th}$  time point

$x_{ij}$ : expression of mRNA <sub>$j$</sub>  on  $i^{th}$  time point

$\epsilon_i$ : error term

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_m = 0 \quad (2)$$

$\beta_1 = \dots = \beta_m$  are sample from common distribution with mean 0 and variance  $\tau^2$

$$H_0 : \tau^2 = 0 \quad (3)$$

*A global test for groups of genes: testing association with a clinical outcome.*  
J.J. Goeman, *et al.* 2004, Bioinformatics.

## 3 Preprocessing

### 3.1 miRNA data

Input PEER residuals (z-transformed) from files  
`EUR363.MirnaQuantCount.1.2N.50F.samplename.resk10.norm.txt` and  
`YRI89.MirnaQuantCount.1.2N.50F.samplename.resk5.norm.txt`. Total number of samples is 449 with 705 microRNAs.

### 3.2 mRNA

`GD662.TrQuantCount.45N.txt` file contains 183085 rows (transcripts or mRNAs) and columns 666 (samples), excluding annotation.

Using the file `GD667_mRNA.SampleInformation_270712.txt` 462 samples were selected.

mRNAs that have non-zero count in at least 90% of the samples are retained. This reduces from 183085 to 18193 mRNAs for further analysis.

Convert Ensembl gene identifiers to Entrez gene identifiers gave 15442 mapped transcripts of which 8505 were unique.

Covert Ensembl transcript identifiers to Entrez gene identifiers gave 15298 mapped transcripts of which 8479 were unique.

When multiple Entrez gene identifiers exist, one Ensembl transcript identifier, the first was used, arbitrarily. Furthermore, only one transcript (unique Entrez identifier) was used. This resulted in an almost 90% reduction of the data.

Raw mRNA counts are corrected using the RLE algorithm of `edgeR`.

## 4 Integrated Analysis

Matched miRNA and mRNA samples were used.

Table 1: Top 20 significant microRNAs

microRNAs	P-value	# Targets
hsa-miR-2277-5p	8.05e-04	4
hsa-miR-192-5p	5.45e-03	93
hsa-miR-877-5p	5.61e-03	71
hsa-miR-106b-5p	8.58e-03	709
hsa-miR-548o-3p	9.84e-03	401
hsa-miR-548o-2-3p	1.00e-02	401
hsa-miR-141-3p	1.18e-02	417
hsa-miR-423-3p	1.30e-02	2
hsa-miR-1243-5p	1.32e-02	85
hsa-miR-4524a-5p	1.43e-02	326
hsa-miR-16-2-5p	1.47e-02	741
hsa-miR-16-1-5p	1.49e-02	741
hsa-miR-26b-3p	1.63e-02	536
hsa-miR-491-5p	2.19e-02	89
hsa-miR-129-1-5p	2.34e-02	302
hsa-miR-129-2-5p	2.34e-02	302
hsa-miR-181d-5p	2.65e-02	664
hsa-miR-363-3p	2.89e-02	492
hsa-miR-30d-5p	2.93e-02	769
hsa-miR-378i-5p	2.99e-02	107

## 5 Future Plans

Integrated analysis on 3' UTR exon-level using microRNA Family (conserved seed). TargetScan lift from Marc Friedleander.

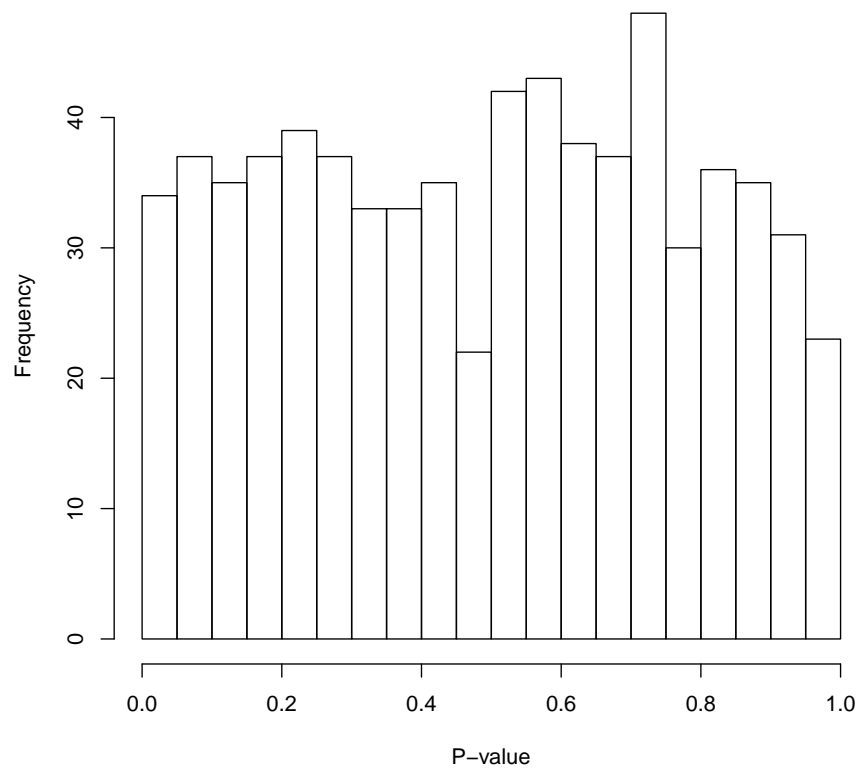


Figure 1: Histogram of P-values from the integrated analysis.

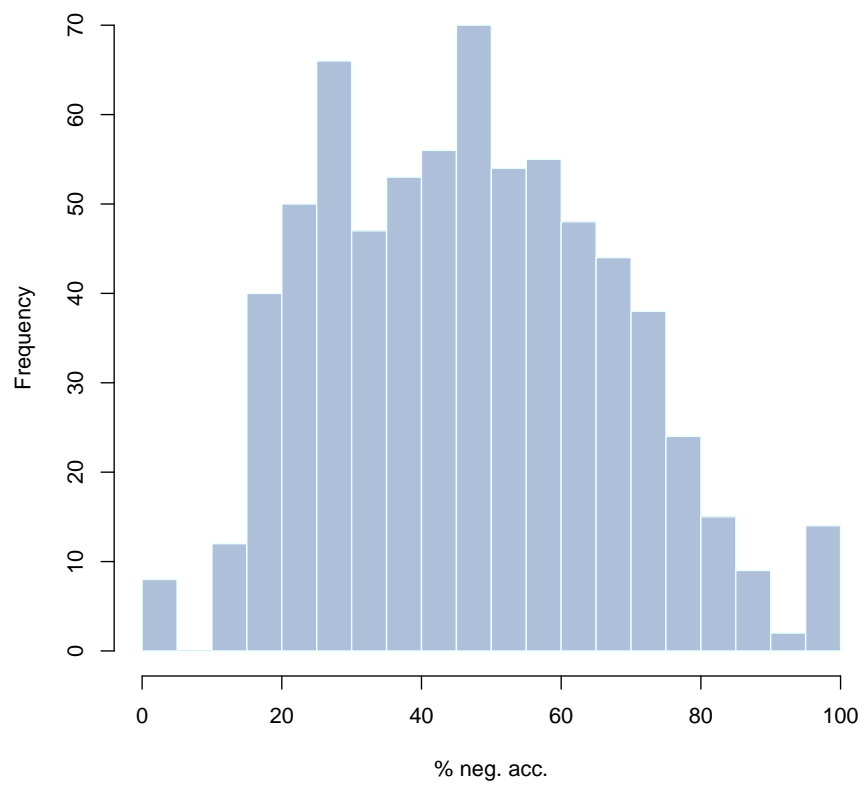


Figure 2: Distribution of the number of negatively associated targets.

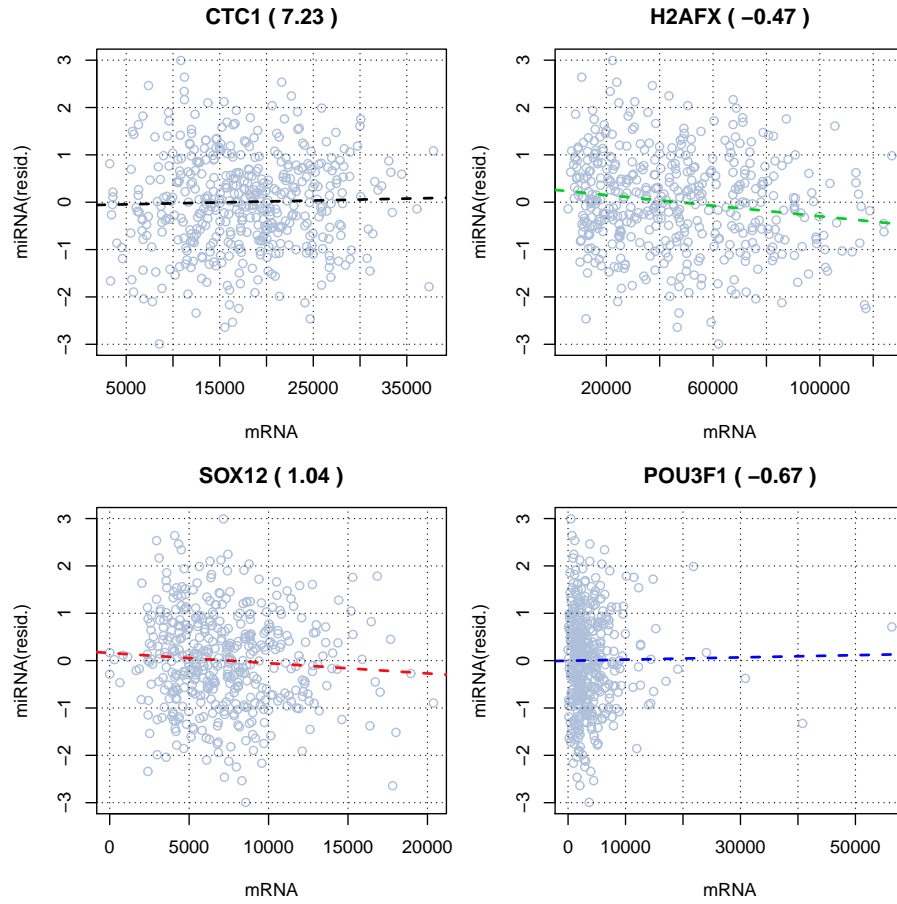


Figure 3: Most significant microRNA hsa-miR-2277-5p with the four predicted targets. Within brackets the z-score or contribution of the target to the overall significance of the microRNA.